Universidade de Brasília
IE - Departamento de Estatística

Estágio Supervisionado 2

Modelos Lineares Mistos

Aplicado na modelagem de níveis de glicemia

Gustavo Juntolli Vilhena

Brasília

Maio de 2013

Universidade de Brasília
IE - Departamento de Estatística

Estágio Supervisionado 2

Modelos Lineares Mistos

Aplicado na modelagem de níveis de glicemia

Gustavo Juntolli Vilhena

Relatório Final

Orientador: Prof.º Eduardo Monteiro de Castro Gomes

Brasília

Maio de 2013

Sumário

Lista de Figuras

Resumo

Aborda-se no trabalho a escolha de um modelo linear misto, sendo este, ajustado aos dados oriundos do Projeto Doce DESAFIO, que possuem medidas repetidas. Dessa forma, envolve-se a escolha dos efeitos aleatórios, dos efeitos fixos e estrutura da matriz de covariâncias. Para isso, foram utilizadas técnicas gráficas e analíticas. Sobre tudo, teste para significância dos efeitos fixos e critérios de informação, AIC e BIC. Utilizou-se, para o desenvolvimento proposto, o *software* livre R com sua função lme().

Palavras-chave: Modelo linear misto; Estrutura de covariância; Análise de resíduos; Critérios de informação.

Abstract

This study focuses on the job choosing a linear mixed model, which is adjusted to the data from the *Projeto Doce DESAFIO* that have repeated measures. Thus, engages the choice of random effects, fixed effects and structure of the covariance matrix. For this, we used graphical and analytical techniques. Above all, test for significance of fixed effects and information criteria, AIC and BIC. It was used for the proposed development, free software R with its function lme ().

Keywords: Linear mixed model; Covariance structure; Residue analysis; Criteria information.

1 Introdução

Atualmente, com a correria do dia-a-dia, tem sido cada vez mais difícil, para a população seguir uma rotina de exercícios e ainda manter uma alimentação saudável. Tendo em vista também, que a automação se torna mais presente na vida das pessoas, percebe-se um crescente aumento nos índices de obesos, que geralmente pode-se associar a diversos tipos de doença. Segundo pesquisas da Organização Mundial da Saúde (OMS), estatísticas revelam que mais de 115 milhões de pessoas sofrem de problemas relacionados com a obesidade nos países em desenvolvimento. No Brasil, a situação é bem semelhante. Cerca de 40% da população está acima do peso.

Das doenças que podem ser causadas pela obesidade, destaca-se a diabetes que consiste no aumento crônico da glicemia, alteração do metabolismo de gorduras, proteínas e carboidratos. Tornando-se evidente o crescimento da parcela populacional doente.

Porém, sabe-se que o ganho de peso não é o único fator gerador da doença, existe também o genético, que entre outros fatores, possui grande importância no aparecimento da diabetes. É possível, com alimentação correta e atividades físicas regulares atrasar o aparecimento ou o agravamento da doença.

Foi baseando-se nisso que várias medidas vêm sendo tomadas, entre elas criação, pela OMS, da Federação Internacional de Diabetes (IDF). Programas de educação no assunto foram criados, entre eles o Doce DESAFIO, com um forte apelo para realização de exercícios físicos orientados e exames clínicos, para avaliar níveis de glicemia, pressão arterial e frequência cardíaca, dentre outros.

O estudo, quando ligado à área de saúde, já possui grande importância para a população, esse em especial, será relacionado à análise de questões oriundas da diabetes. Obtendo-se os dados do projeto, serão feitas modelagens dos níveis de glicemia, o que traz a possibilidade de verificação da existência de variáveis que possam influir no controle da doença. E com isso, aumentar a qualidade de vida.

2 Referencial Teórico

A regressão é a técnica utilizada para estimar uma relação que possa existir na população. A análise de regressão compreende a análise de dados amostrais para saber, se e, como duas ou mais variáveis estão relacionadas uma com a outra numa população.

A regressão linear simples constitui uma tentativa de estabelecer uma equação matemática linear, que descreva o relacionamento entre duas variáveis. Há diversas formas de utilização de equações de regressão:

- Estimar valores de uma variável, com base em valores conhecidos da outra;
- Explicar valores de uma variável em termos da outra, ou seja, confirmar uma relação de causa e efeito entre duas variáveis;
- Predizer valores futuros de uma variável.

Em diferentes situações, há interesse em se estudar o comportamento de alguma característica, de uma variável resposta dos elementos para uma ou mais populações, ao longo de uma condição de avaliação ou escala ordenada, como peso ou tempo.

Segundo Davidian (2007) dados longitudinais são dados na forma de medições repetidas na mesma unidade (humana, vegetal, lote, amostra, etc.) ao longo do tempo. Os dados são coletados rotineiramente, possuindo uma ampla gama de aplicações, incluindo a agricultura e as ciências da vida, a pesquisa médica e de saúde pública e ciências físicas e engenharia.

Estudos longitudinais, segundo LIANG e ZEGER (1986), são de particular interesse quando o objetivo é avaliar variações globais ou individuais ao longo do tempo. Finalmente, alguns parâmetros de interesse podem ser estimados de forma mais eficiente com a utilização de planejamentos longitudinais do que sob outros tipos de planejamentos com o mesmo número de observações.

Lima (1996) relatou que, para a análise dos dados de um experimento com medidas repetidas (dados longitudinais), existem técnicas utilizadas que vão desde análise de variância uni e multivariada, até a metodologia baseada em modelos lineares mistos, com modelagem da estrutura da matriz de variâncias e covariâncias. Everitt (2004) destacou a importância do uso desses modelos na analise de dados com as medidas repetidas.

2.1 Modelos Lineares Mistos

2.1.1 Introdução

Há várias abordagens para o modelo linear misto, como modelo linear clássico, modelo de componentes de variâncias e também modelos hierárquicos multiníveis (NATIS, 2000), entre outras alternativas de análise de dados com medidas repetidas como modelos lineares generalizados mistos (Silvano, 2003).

Segundo Pinheiro (1994), com a possibilidade de modelar a correlação intra-indivíduos – muitas vezes presente em dados agrupados, tornou os modelos lineares mistos em um tema crescente na estatística nas últimas décadas. Observações realizadas, sob um mesmo indivíduo, tendem a ser correlacionadas, e os modelos lineares mistos constituem uma ferramenta conveniente para modelar essa dependência

intra-indivíduos.

2.1.2 Especificação do modelo

Os modelos lineares mistos permitem a utilização de várias estruturas de componentes de covariâncias, que foram tratados primeiramente por Laird e Ware (1982), consideraram, no primeiro estágio, os efeitos fixos para obtenção da curva polinomial média. Permitindo, no segundo estágio, diferentes curvas para cada indivíduo.

O modelo linear misto é expresso na seguinte forma matricial:

$$\mathbf{Y_i} = \mathbf{X_i}\beta + \mathbf{Z_i}\mathbf{b_i} + \epsilon_{\mathbf{i}}, \mathrm{i} = 1, ..., \mathrm{N} \tag{1}$$

Onde $\mathbf{y_i}$ representa um vetor$(n_i * 1)$ de respostas da i-ésima unidade experimental ou indivíduo, $\mathbf{X_i}$ é uma matriz$(n_i * p)$ de especificação (conhecida e de posto completo) dos efeitos fixos, β é um vetor $(p * 1)$ de parâmetros de (efeitos fixos), $\mathbf{Z_i}$ é uma matriz$(n_i * q)$de especificação (conhecida e de posto completo) dos efeitos aleatórios, $\mathbf{b_i}$ é um vetor $(q * 1)$ de efeitos aleatórios com vetor de média 0 e matriz de covariância $\mathbf{D}$ e $\epsilon_{\mathbf{i}}$ é um vetor $(n_i * 1)$ com erros aleatórios com vetor de médias 0 e matriz de covariância $cov(\epsilon) = R_i(\lambda)$, positiva definida, com λ um vetor de parâmetros desconhecidos. Há N unidades experimentais e n_i observações feita na i-ésima unidade experimental.

A forma de especificação da matriz de $\mathbf{X_i}$ é bastante similar àquela utilizada nos modelos de regressão. Suas colunas podem estar associadas:

- Aos fatores que definem a estrutura das subpopulações (grupos ou tratamentos);

- Ao fator tempo, identificando, por exemplo, a forma da curva a ser ajustada;
- Às covariáveis, cujos efeitos na resposta média deseja-se pesquisar.

Estes pressupostos implicam que a matriz de covariância de $cov(y_i) = \sum = Z_i D Z_i^T + R_i$. Esse método de estruturar a matriz de covariâncias $\sum i$ tem como atrativo a possibilidade de:

- Englobar as abordagens uni e multivariada que são comumente utilizadas na análise de dados longitudinais;
- Lidar com dados perdidos, por causa da facilidade de construir a verossimilhança somente dos dados observados;
- Usar estruturas relacionadas com séries temporais ou estruturas mais complexas.

Nas situações em que o objetivo da análise é ajustar curvas (de crescimento), os modelos lineares mistos assumem a existência de curvas subpopulacionais fixadas ($X_i\beta$), havendo variações aleatórias – ($Z_i b_i$) em torno das curvas individuais, e também existem variações aleatórias de medidas (ϵ_i) em torno da curva média.

O modelo linear misto, usualmente, é especificado em termos das respostas condicionadas aos efeitos aleatórios, de modo que assumindo y_i um vetor de medidas repetidas para o i-ésimo indivíduo, satisfaça:

$$y_i | b_i \sim N(X_i\beta + Z_i b_i, R_i)$$

$$b_i \sim N(0, D)$$

2.2 Estrutura das Matrizes de Covariância

As estimativas e os erros padrões de efeitos fixos, diagnósticos e inferências são afetados de acordo com a forma da estrutura de covariância escolhida, a depender da informação empírica da estrutura dos dados e muitas vezes da disponibilidade computacional.

A estrutura dos modelos lineares mistos permite a consideração de matrizes especiais de covariância, buscando representar de forma mais precisa a variabilidade dos dados, ou seja, consegue absorver as diferentes informações, levando em consideração se os dados são independentes, dependentes, correlacionados etc.

Serão apresentadas, a seguir, algumas estruturas da matriz $\mathbf{D}$ e $\mathbf{R_i}$:

1. Matriz de componentes de variância (VC) – impõe variâncias iguais nas n_i observações e adota independência entre elas:

$$\begin{pmatrix} \sigma^2 & 0 & 0 & 0 \\ 0 & \sigma^2 & 0 & 0 \\ 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & \sigma^2 \end{pmatrix}$$

2. Matriz de simetria composta (CS) – determina variâncias iguais nas n_i observações e mesma covariância entre medidas feitas em ocasiões distintas:

$$\begin{pmatrix} \sigma^2 & \sigma_1 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma^2 & \sigma_1 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_1 & \sigma_1 & \sigma_1 & \sigma^2 \end{pmatrix}$$

3. Simetria Composta Heterogênea

$$\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho & \sigma_1\sigma_4\rho \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho \\ \sigma_3\sigma_1\rho & \sigma_3\sigma_2\rho & \sigma_3^2 & \sigma_3\sigma_4\rho \\ \sigma_4\sigma_1\rho & \sigma_4\sigma_2\rho & \sigma_4\sigma_3\rho & \sigma_4^2 \end{pmatrix}$$

Adota parâmetros de variâncias diferentes para cada elemento da diagonal principal, obtendo a raiz quadrada desses elementos nos elementos fora da diagonal principal sendo ρ a correlação entre as medições, satisfazendo $|\rho| < 1$. Possui $n_i + 1$ parâmetros.

4. Não Estruturada

$$\begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_2^2 & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_3^2 & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_4^2 \end{pmatrix}$$

Variâncias distintas para cada uma das n_i ocasiões e covariâncias diferentes entre medidas feitas em ocasiões diferentes. Envolvendo $n_i(n_i+1)/2$ parâmetros.

5. Estrutura AR(1)

$$\begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{pmatrix}$$

Denota variâncias iguais nas diversas ocasiões e correlação decrescente, com o aumento do intervalo entre as ocasiões, apenas dois parâmetros são considerados.

6. Estrutura ARH(1)

$$\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho & \sigma_1\sigma_3\rho^2 & \sigma_1\sigma_4\rho^3 \\ \sigma_2\sigma_1\rho & \sigma_2^2 & \sigma_2\sigma_3\rho & \sigma_2\sigma_4\rho^2 \\ \sigma_3\sigma_1\rho^2 & \sigma_3\sigma_2\rho & \sigma_3^2 & \sigma_3\sigma_4\rho \\ \sigma_4\sigma_1\rho^3 & \sigma_4\sigma_2\rho^2 & \sigma_4\sigma_3\rho & \sigma_4^2 \end{pmatrix}$$

É a generalização da AR(1), impondo variâncias e covariâncias diferentes. Envolve $n_i + 1$ parâmetros.

7. Estrutura Toeplitz

$$\begin{pmatrix} \sigma^2 & \sigma_1 & \sigma_2 & \sigma_3 \\ \sigma_1 & \sigma^2 & \sigma_1 & \sigma_2 \\ \sigma_2 & \sigma_1 & \sigma^2 & \sigma_1 \\ \sigma_3 & \sigma_2 & \sigma_1 & \sigma^2 \end{pmatrix}$$

A estrutura de covariâncias de um processo de médias móveis de ordem $q = n_i - 1$. Possui n_i parâmetros.

8. Estrutura ARMA(1,1)

$$\begin{pmatrix} \sigma^2 & \sigma^2\gamma & \sigma^2\gamma\rho & \sigma^2\gamma\rho^2 \\ \sigma^2\gamma & \sigma^2 & \sigma^2\gamma & \sigma^2\gamma\rho \\ \sigma^2\gamma\rho & \sigma^2\gamma & \sigma^2 & \sigma^2\gamma \\ \sigma^2\gamma\rho^2 & \sigma^2\gamma\rho & \sigma^2\gamma & \sigma^2 \end{pmatrix}$$

Essa estrutura é associada a séries temporais com parâmetros auto-regressivo ρ, componente de média móveis γ, sendo σ^2 a variância residual. Envolve três parâmetros.

9. Estrutura Ante dependencia de ordem 1

$$\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_1 & \sigma_1\sigma_3\rho_1\rho_2 & \sigma_1\sigma_4\rho_1\rho_2\rho_3 \\ \sigma_2\sigma_1\rho_1 & \sigma_2^2 & \sigma_2\sigma_3\rho_1 & \sigma_2\sigma_4\rho_2 \\ \sigma_3\sigma_1\rho_1\rho_2 & \sigma_3\sigma_2\rho_2 & \sigma_3^2 & \sigma_3\sigma_4\rho_1 \\ \sigma_4\sigma_1\rho_1\rho_2\rho_3 & \sigma_4\sigma_2\rho_2\rho_3 & \sigma_4\sigma_3\rho_3 & \sigma_4^2 \end{pmatrix}$$

Determina parâmetros de variâncias diferentes para cada elemento da diagonal, sendo os elementos fora da diagonal principal funções de variâcias e do k-ésimo parâmetros de autocorrelação, satisfazendo $|\rho_k| < 1$. Esta estrutura permite que as variâncias sejam diferentes e é aplicavel em estudos longitudinais em que as condições de avaliação não são igualmente espaçadas, apresentam heterogeneidade de variância e correlação serial. Possui $2n_i - 1$ parâmetros.

10. Estrutura Toeplitz Heterogênea

$$\begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_1 & \sigma_1\sigma_3\rho_2 & \sigma_1\sigma_4\rho_3 \\ \sigma_2\sigma_1\rho_1 & \sigma_2^2 & \sigma_2\sigma_3\rho_1 & \sigma_2\sigma_4\rho_2 \\ \sigma_3\sigma_1\rho_2 & \sigma_3\sigma_2\rho_1 & \sigma_3^2 & \sigma_3\sigma_4\rho_1 \\ \sigma_4\sigma_1\rho_3 & \sigma_4\sigma_2\rho_2 & \sigma_4\sigma_3\rho_1 & \sigma_4^2 \end{pmatrix}$$

Também associada a dados de séries temporais, igualmente espaçados, com parâmetros de variâcias diferentes para cada elemento da diagonal, sendo

os elemenos fora da diagonal principal, funções de variâncias e do k-ésimo parâmetro de autocorrelação $|\rho_k| < 1$.

2.3 Estimação dos parâmetros

No caso do modelo linear misto, dentre vários outros métodos existentes, além de refinamentos e novos métodos que são apresentados a cada dia, utiliza-se em sua maioria, basicamente dois: o método de Máxima Verossimilhança e o de Máxima Verossimilhança Restrita, os quais serão apresentados a seguir.

2.3.1 Máxima Verossimilhança

Como já é sabido, se utiliza o método de Máxima Verossimilhança - MV para a estimação dos parâmetros nos modelos lineares usuais, onde há normalidade dos erros, e também em modelos lineares generalizados. O método se baseia no comportamento probabilístico do erro associado, no caso dos modelos lineares usuais. Ou no comportamento da variável de interesse nos modelos lineares generalizados.

Nos modelos lineares mistos, a estimação dos parâmetros não se dá da mesma forma, é diferenciada quando comparada a outras estimações. Para os modelos usuais, a Máxima Verossimilhança, aplica-se de forma direta, tendo em vista, que o comportamento das partes pertencentes ao modelo e a matriz de covariáveis é fixa ou observável. Não sendo possível, aplicá-la na modelagem mista, pois não se conhece o comportamento probabilístico dos efeitos aleatórios.

Com isso, a inferência relacionada aos modelos lineares mistos é fundamentada na densidade marginal do vetor de resposta Y_i. Dessa forma, Verbeke e Molembergs (2000) apresentam a forma clássica de estimação baseada na maximização da função

verossimilhança marginal:

$$L_{MV}(\theta) = \prod_{i=1}^{N} \{(2\pi)^{-n_i/2} |\Sigma_i(\alpha)|^{-1/2} \exp(-1/2(y_i - X_i\beta)'(\Sigma_i^{-1}(\alpha))(y_i - X_i\beta))\}$$

No qual α denota o vetor com todos os parâmetros de varâncias e covariâncias (usualmente chamados de componentes de variância) encontrado em $\sum_i = Z_i D Z_i^T + R_i$, que consiste de $n_i(n_i+1)/2$ elementos diferentes em D e de todos os parâmetros em R_i. Seja θ o vetor de parâmetros para o modelo marginal y_i.

O estimador de máxima verossimilhança de β, obtidos a partir de maximização da função, condicionada a α é dada por (LAIRD; WARE, 1982):

$$\widehat{\beta}(\alpha) = (\sum_{i=1}^{N} (X_i' \Sigma_i^{-1} X_i))^{-1} \sum_{i=1}^{N} (X_i' \Sigma_i^{-1} y_i)$$

2.3.2 Máxima Verossimilhança Restrita

Este método, também retratado como Máxima Verossimilhança Residual - MVR, e proposto por Patterson and Thompson (1971), tem como objetivo estimar os componentes da variância. Se, para isso, fosse utilizado a Máxima Verossimilhança, a estimação de σ^2 seria viesada, subestimando as estimativas dos parâmetros. Segundo Davis (2002), o método MRV está associado ao "contraste de erros" e não às observações reais, pois, dessa maneira, as estimativas para os componentes da variância são estimados com menos viés. Uma definição de estimação, fornecendo uma conveniente forma computacional é:

$$L_{MVR}(\theta) = |\sum_{i=1}^{N} (X_i' \Sigma_i^{-1} X_i)|^{-1/2} L_{MV}(\theta)$$

Para a maximização do logaritmo da verossimilhança restrita são necessários métodos iterativos, como o método de Newton-Raphson, método Fisher scoring,

sendo o primeiro, com as modificações propostas por Jennrich e Schluchter(1986), considerado o melhor, em relação ao tempo total para atigir a convergência.

2.4 Escolha do modelo

Na literatura existem critérios que, para escolha do modelo, fornecem auxílios na decisão de qual modelo selecionar. A seleção do melhor modelo não significa apenas verificar a melhor forma para se estruturar a parte fixa (efeitos fixos) e covariância. É necessário também, averiguar e identificar os efeitos aleatórios, bem como, a estimação e comparação entres os modelos.

Dessa forma, torna-se simples a avaliação e seleção do modelo, tendo em vista que existem alguns métodos para selecionar o modelo mais adequado e, com isso, se ajustando melhor aos dados. Rocha (2004) propõe técnicas gráficas e analíticas, auxiliando na escolha das matrizes dos modelos lineares mistos, pois, em estudos longitudinais, é razoável utilizar informações que remetem o comportamento da resposta durante as avaliações na modelagem da estrutura de covariância intra-unidades da amostra.

Wolfinger (1993) apresenta técnicas que tornam úteis uma coleção de estruturas de covariância para dados de medidas repetidas, ampliando significativamente a disponibilidade de um conjunto de modelos estatísticos para a explicação da variabilidade dos dados.

A partir dessa estruturação para uma boa análise dos dados, Pinheiros e Bates (2000) ressaltaram um crescimento na popularização da modelagem com a utilização de modelos lineares mistos, sendo explicada pela maior flexibilidade oferecida na

estruturação da correlação intra-indivíduos e pela maior confiabilidade e disponibilidade de SOFTWARE, utilizado no ajuste.

Sendo assim, faz-se o uso, com maior frequência, de dois critérios: o AIC (Akaike Information Criterion) e o BIC (Bayesian Information Criterion). Sendo eles definidos como:

$$AIC = -2l + 2p$$

$$BIC = -2l + p\log(n)$$

Em que l representa o máximo da log-verossimilhança, p a quantidade de parâmetros e n a quantidade de observações. Assim, adota-se, como decisão, o melhor modelo o que apresentar o menor valor, valendo tanto para o AIC, como também, para o BIC. É importante ressaltar que, entre os dois, o BIC possui uma consistência maior. Outra característica relacionada ao BIC é a de maior penalização aos modelos mais complexos, além de incluir a quantidade de observações em seu cálculo. Então, sugere Gurka (2006), quando se utiliza o método de estimação MVR, o uso do BIC para tomar decisões quando há divergência no modelo indicado pelos métodos.

Bates (2000) informa que quando se ajusta o modelo utilizando a máxima verossimilhança os valores de AIC e BIC podem ser comparados entre quaisquer modelos ajustados para os mesmos dados. Entretanto ao se utilizar a máxima verossimilhança restrita, seus valores, e incluindo o log-verossimilhança, somente podem ser comparados entre modelos que possuam a mesma estrutura de efeitos fixos.

2.5 Testes de hipóteses

Independentemente do modelo adotado, os testes de hipóteses são uma parte fundamental no processo de ajuste. Sendo eles responsáveis pela determinação da significância do modelo e das estimativas dos parâmetros nele envolvios. Em modelos lineares mistos os testes de hipóteses são aproximados. No entanto, são vários os testes descritos na literatura, tais como Wald, Escore e o teste da Razão de Verossimilhança.

2.5.1 Teste da razão de verossimilhança

Para modelos ajustados pelo método de máxima verossimilhança, o teste mais utilizado, comumente, é o teste da Razão de Verossimilhança segundo PINHEIROS e BATES (2000). Tal teste baseia-se na razão entre as Verossimilhanças dos dois modelos ajustados, sendo fundamental sua importância para a verificação se um modelo que possui número menor de parâmetros se ajusta de forma tão boa quanto o modelo com a quantidade total de parâmetros.

A estatística do teste da Razão de Verossimilhança é dada por:

$$\xi_{RV} = 2[\log(L_2) - \log(L_1)]$$

O teste apresenta distribuição quiquadrado com r graus de liberdade, em que r é a diferença entre a quantidade de parâmetros dos modelos testados, onde L_1 é o valor maximizado da log-verossimilhança do modelo reduzido e l_2 do modelo completo. Pinheiro e Bates (2000) não recomendam utilizar a Razão de Verossimilhanças para testar efeitos fixos, uma vez que o teste segue uma distribuição de referência qui-quadrado e assim tende a ser anticonservativo. Além disso, não é indicada a

utilização do teste com o interesse em verificar hipóteses que se remetem aos efeitos fixos, quando utilizada a MVR, uma vez que, ao utilizar tal método, os efeitos fixos são desconsiderados.

A solução alternativa sugerida por Pinheiro e Bates (2000) é condicionar a especificação desses efeitos às estimativas das variâncias e covariâncias dos efeitos aleatórios. Este teste condicional é dado pelos teste-F e teste-t usuais, como definidos nos modelos lineares, sendo condicionados à

$$\sigma_R^2(\omega) = s^2 = \frac{RSS}{M-p}$$

onde RSS é a soma de quadrados do resíduo, ω refere-se aos parâmetros envolvidos nos efeitos fixos, M é a soma dos n_i e p a quantidade de parâmetros.

2.5.2 Teste de Wald

O teste de Wald é utilizado para avaliar a significância dos efeitos fixos do modelo(1). A estatística de Wald para testar $H_0 : C\beta = 0$, sendo $C_{(c*p)}$ uma matriz de constantes conhecidas e de posto completo ($c \leq p$) escrita como:

$$Q_c = (C\widehat{\beta})'[C\widehat{cov}(\widehat{\beta})C']^{-1}(C\widehat{\beta})$$

em que $\widehat{cov}(\widehat{\beta})$ é uma estimativa da matriz de covariâncias de $\widehat{\beta}$. A estatística Q_c tem distribuição assintótica *quiquadrado* com c graus de liberdade, sob H_0, ao dividir Q_c por c, é obtido uma nova estatística com distribuição F com c e p-posto(X) graus de liberdade.

Verbeke e Molenberghs (2000) criticam a adequação do teste de Wald, quando utilizado em modelos lineares mistos, que são especificados condicionalmente aos

efeitos aleatórios. Pois o teste não leva em conta a estimativa dos parâmetros de efeito aleatório, podendo, então, subestimar a variação dos efeitos fixos.

2.6 Previsão

Um dos principais interesses no ajuste de um modelo é a previsão. Como é definido por Pinheiros e Bates (2000), os valores ajustados são as previsões realizadas para as respostas observadas. Sendo elas de grande importância na verificação de quão adequado é o modelo, de tal forma que deseja-se previsões o mais próximo possível, determinado menores magnitudes dos resíduos.

Segundo Pinheiro e Bates (2000), a partir do modelo linear misto existe a possibilidade de se fazer previsões em dois níveis diferentes: o nível populacional e o nível individual. No nível populacional, os valores ajustados representam os valores preditos para os valores marginais esperados da variável resposta, enquanto que no nível individual, as predições representam os valores esperados condicionados aos efeitos aleatórios estimados daquele indivíduo.

$$E[y_i] = x_i'\beta$$

$$E[y_i|b_i] = x_i'\beta + z_i'b_i$$

sendo x_h o vetor de covariáveis associado aos efeitos fixos e $z_h(i)$ vetor de covariáveis associadas aos efeitos aleatórios do i-ésimo grupo. Chegando aos seguintes valores preditos:

$$\widehat{y_i} = x_i'\widehat{\beta}$$

$$\widehat{y_i} = x_i'\widehat{\beta} + z_i'\widehat{b_i}$$

onde $\widehat{\beta}$ e $\widehat{b}$ são o BLUE e o BLUP, respectivamente.

2.7 Análise de resíduos

No que tange o estudo da adequacidade do ajuste de modelo, seja ele de qualquer natureza - modelos lineares em sua forma mais simples à modelos generalizados lineares e não-lineares, como também modelos mais complexos, é de importância indiscutível realizar a análise de resíduos.

Assim, o estudo da adequabilidade de forma geral, visa a verificação das suposições impostas pelo modelo, entretanto tal estudo vai além da verificação de suposições, tendo como preocupação, também, verificar a forma como as observações influenciam no ajuste do modelo. É importante ressaltar que cada modelo possui uma determinada estrutura, apesar da metodologia de ajuste - estimação dos parâmetros que geralmente segue o caminho da máxima verossimilhança. A abordagem dos resíduos deve ser cuidadosa, tendo em vista que a estrutura dos resíduos que melhor se encaixa ao estudo da adequacidade varia de modelo para modelo.

Três tipos de erros para os modelos lineares mistos são apresentados por Nobre e Singer (2007). As três abordagens são necessárias para o estudo da adequabilidade devido às suas características, posibilitando estudar um conjunto de diferentes suposições. Sendo eles denominados e dados por:

Erros condicionais

$$\varepsilon = Y - X\beta - Zb$$

Efeitos aleatórios

$$Z_b = E[Y|b] - E[Y]$$

Efeitos marginais

$$\xi = Y - X\beta = Z_b + \varepsilon$$

Segundo Pinheiro e Bates (2000), antes de quaisquer inferências, duas suposições devem ser verificadas nos modelos lineares mistos: se os erros intra-grupos são independentes e identicamente distribuídos seguindo uma distribuição normal com média zero e variância σ^2 e se são independentes dos efeitos aleatórios. A outra suposição refere-se à normalidade dos efeitos aleatórios e são independentes para diferentes grupos.

Pinheiro e Bates (2000) propõem o uso do gráfico de probabilidade normal dos resíduos condicionais para avaliar a suposição de normalidade e o gráfico dos resíduos condicionais versus os valores ajustados para avaliar a suposição de homocedasticidade.Além disso os resíduos condicionais também podem ser utilizados para identificação de pontos discrepantes. Contudo, Nobre(2004), com base na possibilidade dos elementos de $\widehat{\varepsilon}$ apresentarem variâncias diferentes, propõe uma padronização dos resíduos condicionais.

3 Aplicação

A partir da metodologia apresentada, seguirão os resultados obtidos neste capítulo. Sendo assim, será estruturado em quatro partes: descrição dos dados; análise descritiva; ajuste dos dados ao modelo linear misto e, por último, a análise dos resíduos resultantes da modelagem. Para isso, foi utilizado o programa R 2.15.2. E todas as análises foram baseadas com nível de significância de 5%, ou seja, para os testes de hipóteses $\alpha = 0,05$.

3.1 Descrição dos dados

O projeto – Diabetes, Educação em Saúde e Atividades Físicas Orientadas (Doce DESAFIO) – foi instituído pela Faculdade de Educação Física da UnB (Universidade de Brasília), tendo como objetivo a educação em diabetes tornando a prática de atividades físicas o seu pilar, aliado aos medicamentos e boa alimentação. Neste programa participam, gratuitamente, diabéticos de todos os tipos (tipo 1 e tipo 2), pessoas de qualquer idade, e são atendidos participantes de todo DF e região.

As atividades ocorrem duas vezes na semana, em diferentes localizações – Centro Olímpico (UnB), Centro de saúde 4 (Samambaia) e no Centro de Saúde 2 (Sobradinho). Com realização de alongamentos, ginásticas, futebol, musculação, caminhada dentre outros exercícios.

Juntamente com as realizações das atividades, são feitas avaliações de glicemia, pressão arterial, frequência cardíaca, medicação, alimentação e outros dados para que se tenha um bom acompanhamento dos participantes. Sendo oferecidos também, palestras e cursos.

Assim, cada participante é monitorado, obtendo então, valores de seus níveis de glicemia sanguínea quando chega à aula para prática da atividade física, como também, ao final do exercício proposto.

Com isto, é constituído o banco de dados que será utilizado para o estudo sobre a modelagem dos níveis de glicemia, em relação às diversas variáveis pertencentes ao banco de dados. Dentre elas, encontra-se a idade dos participantes, locais de realização da atividade, momento do diagnóstico como diabético, tipo de diabetes, se do tipo 1 ou do tipo 2, os valores glicêmicos iniciais e finais, se é feito o uso de insulina exógena e também a data de ingresso ao programa.

3.2 Análise descritiva

No estudo proposto, serão utilizados os dados armazenados pelo Projeto Doce DESAFIO, sendo assim foi realizado uma analise descritiva para melhor adequação dos dados no estudo do modelo.

Para um quadro total de 38 participantes frequentes, obtém-se que 13 são do sexo masculino e 25 pertencem ao feminino. Quando se analisa o tipo de diabetes, percebe-se que em sua maioria é do tipo 2, 34 usuários do programa, ou seja, a insulina produzida é de baixa eficiência. Dos dados, apenas dois são do tipo 1 – insulina insuficiente, e dois estão no estágio de pré-diabetes. Como é demonstrado nas tabelas seguintes:

Tabela 1: Número de usuários/sexo

Sexo	*Frequencia*
Feminino	25
Masculino	13

Tabela 2: Usuários por tipo de diabetes

TipodeDiabetes	*Frequencia*
Pré-diabetes	2
Tipo I	2
Tipo II	34

Outro ponto importante é verificar se faz aplicação ou não de insulina, sendo que a maioria não faz, correspondendo a 31. A idade média dos participantes do programa é de aproximadamente 62 anos, não possuindo um desvio padrão grande, 12,36. A mediana ficou bem próxima da média – 64,5 anos, o que pode indicar simetria na distribuição das idades. A idade mínima registrada foi de 35 anos e a máxima 85.

Na figura 1: um gráfico do tipo "boxplot" das idades, onde é possível realizar uma comparação entre as idades do sexo feminino com o masculino. Observa-se, então, que as idades medianas são próximas, bem como, as amplitudes das idades do sexo masculino e feminino.

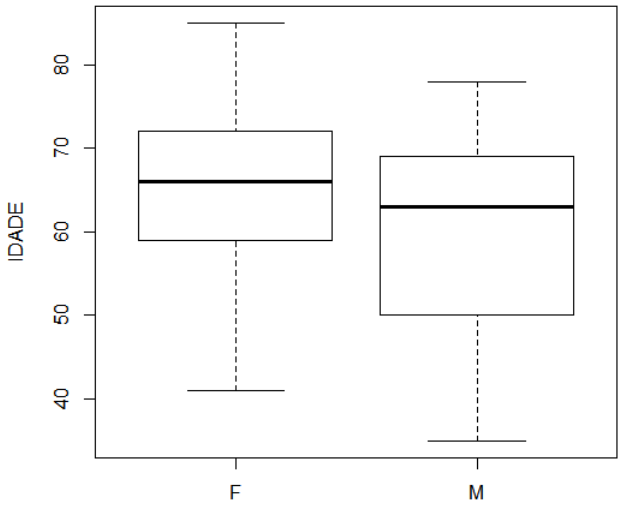

Figura 1: Boxplot das idades/sexo

Dessa forma foi feita, também, uma análise envolvendo o valor médio de níveis de glicemia, assim, para cada aula, obteve-se uma média inicial e final para o grupo de participantes, com isso, pode-se observar que em todos os casos houve redução. A partir desses valores, foi calculada uma média, para as onze aulas, do momento inicial e final, sendo o primeiro igual a 147,4 e o segundo a 116,7. Então, foi gerado outro gráfico do tipo boxplot, figura 2, relacionando níveis de glicemia com o momento da coleta, se inicial ou final.

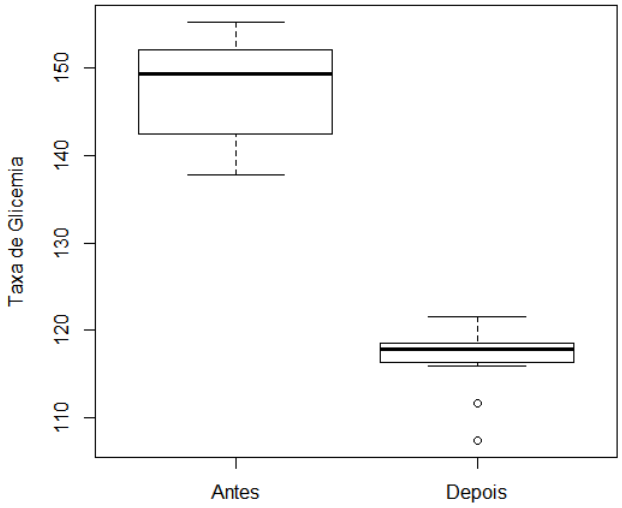

Figura 2: Glicemia por momento da coleta

Pela analise visual dos gráficos, nota-se que, para as médias finais, existem dados discrepantes, para valores mais baixos. Outro ponto a ressaltar é a possibilidade de haver maior controle, nos níveis de glicemia, tendo em vista a menor amplitude do gráfico após a atividade física. Demonstrando uma eficiência, na redução dos níveis glicêmicos, pela prática de exercícios. Com a análise inicial feita, percebe-se que uma forma possível de se trabalhar com os valores da glicemia, tendo em vista, que para

cada individuo, são obtidos os níveis inicial e final por aula, adotou-se a variação da glicemia, ou seja, o valor inicial menos o final para os respectivos alunos e suas aulas.

A partir disso, verifica-se uma possível relação entre a idade do participante com os valores da variação de suas medidas de glicemia, uma melhor visualização pode ser obtida no gráfico de dispersão, figura 3:

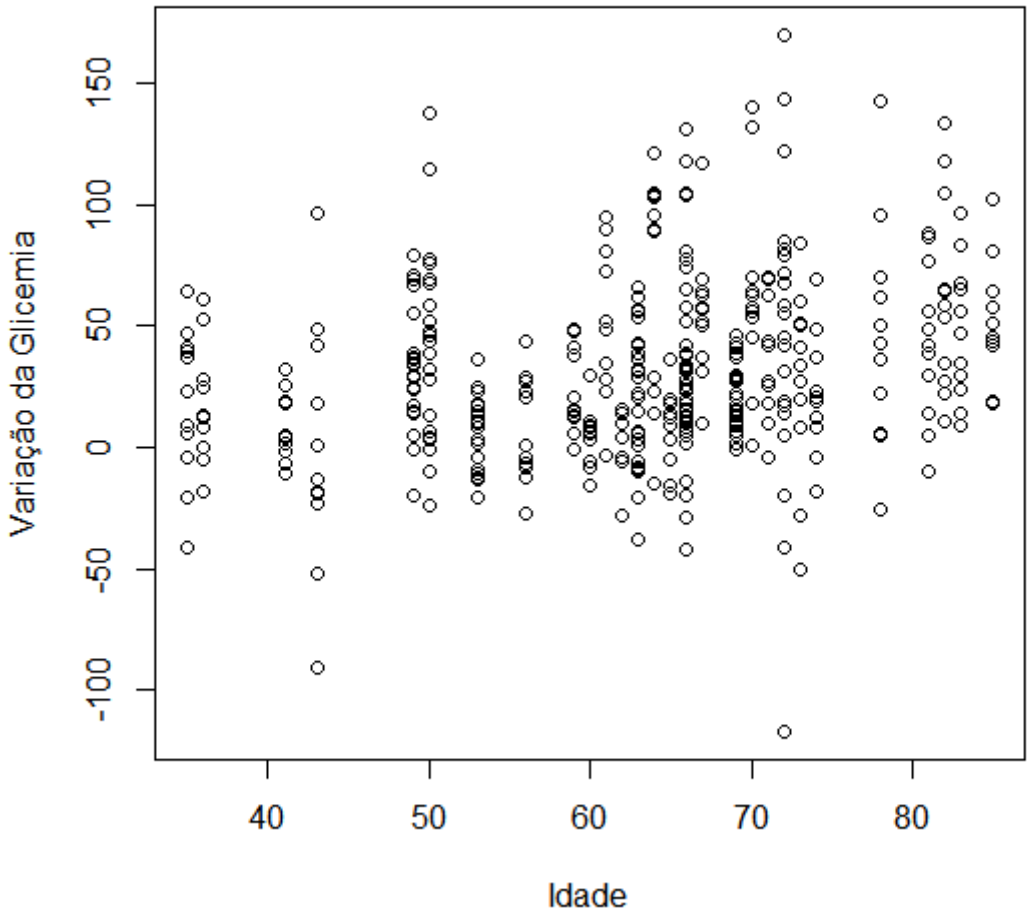

Figura 3: Dispersão da variação glicêmica pela idade

3.3 Motivação

Na área médica é comum a utilização de modelos lineares mistos, tendo em vista, que quando se trata de seres humanos, não há um comportamento padrão entre os diversos indivíduos estudados. Tal comportamento pode ser entendido como uma tendência geral, independente do indivíduo. Porém, quando se faz a análise dos gráficos de perfis, que serão visualizados a seguir, é notória a diferenciação de cada individuo em relação aos valores glicêmicos. E para um bom ajuste, é importante não desprezar essa informação, ou seja, para o estudo proposto existe indicação da utilização do modelo misto. A seguir encontra-se um gráfico com os valores glicêmicos para cada indivíduo:

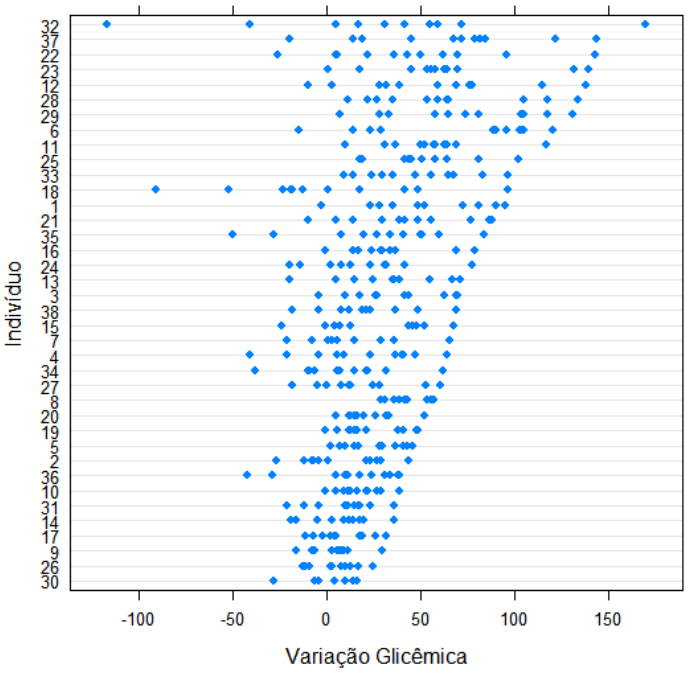

Figura 4: Valores da glicemia para cada indivíduo

Ainda analisando o comportamento dos níveis da glicemia no indivíduo, nota-se que há uma redução quando se observa os valores iniciais e os finais, ou seja, é

provável que a prática da atividade física tenha alguma relação com a mudança nos valores entre o início e o fim de cada aula. Tal característica pode ser vista no gráfico de perfil e, mais intensificada no gráfico de perfil médio, figura 5 e 6 respectivamente:

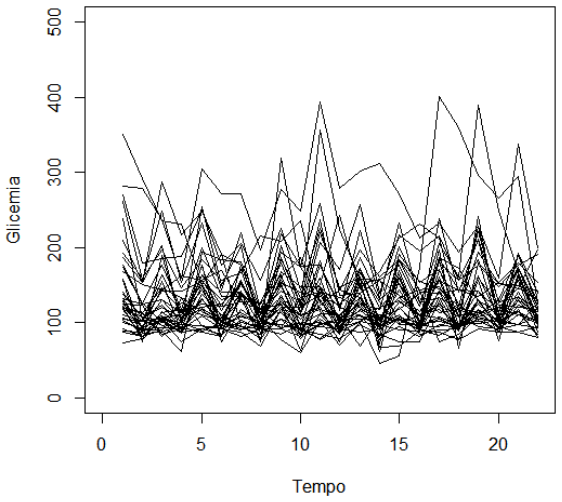

Figura 5: Gráfico de perfil dos participantes

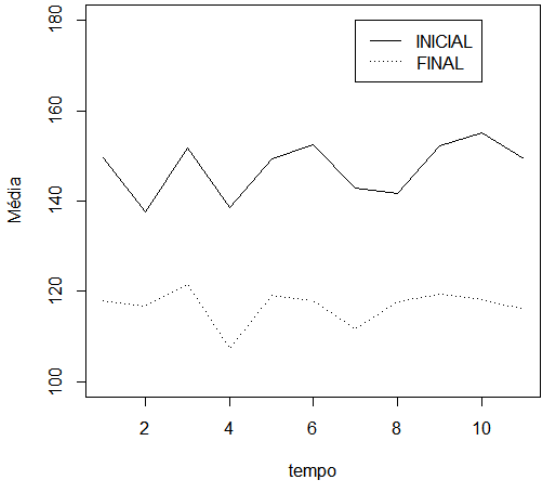

Figura 6: Gráfico do perfil médio

Para melhor compreensão das características de cada participante, individualmente, quanto a variação do nível de glicemia, foi plotado o gráfico de perfil para cada usuário do programa:

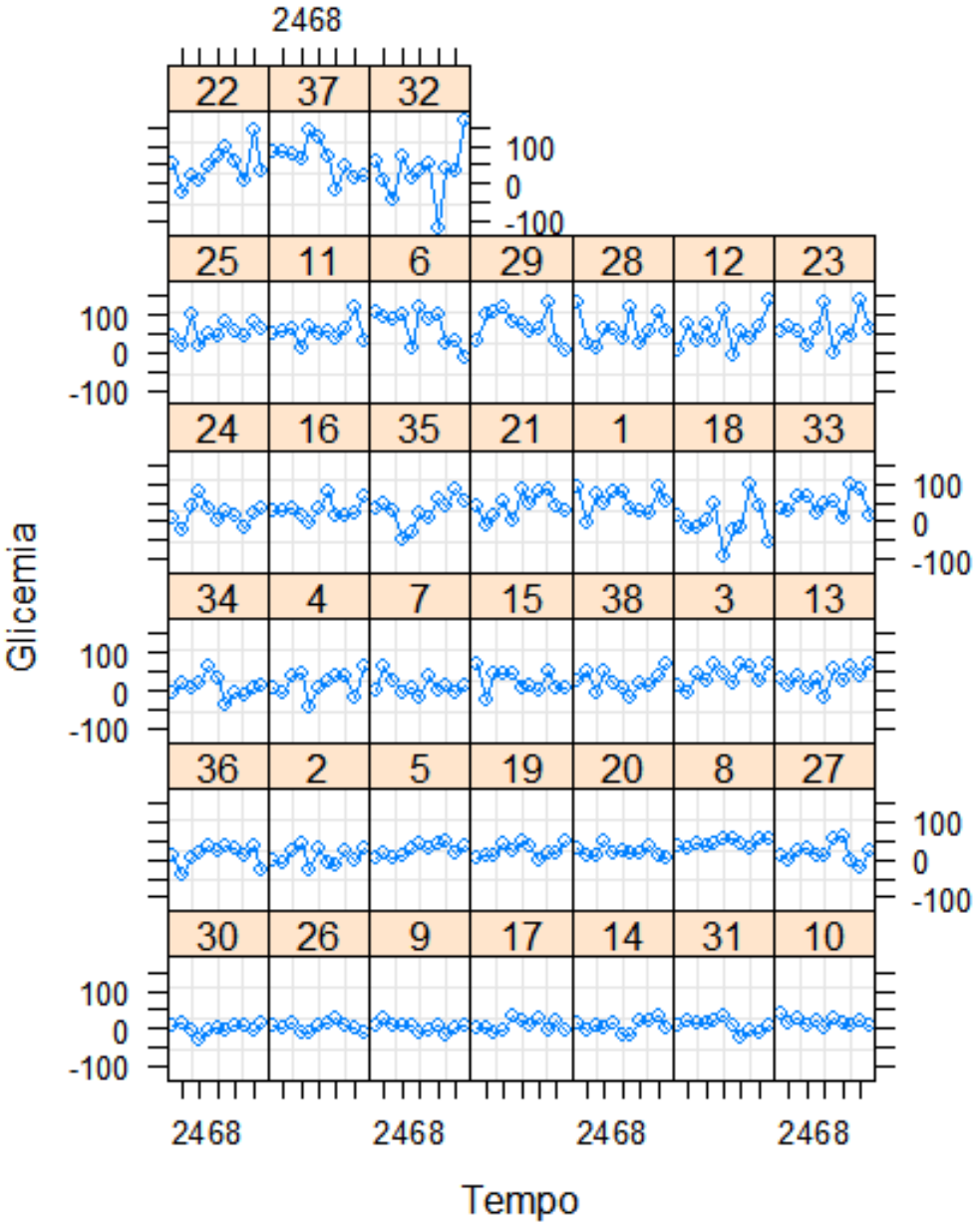

Figura 7: Perfil individual dos participantes

Outro ponto válido, a ser observado é o gráfico do tipo boxplot, apresentado a seguir, dos valores da variação glicêmica para cada tempo e sexo, onde as variações no sexo masculino e feminino, não parecem se diferenciar. O que pode indicar sexo como uma variável fixa não relevante para o modelo, e isto será testado mais a frente.

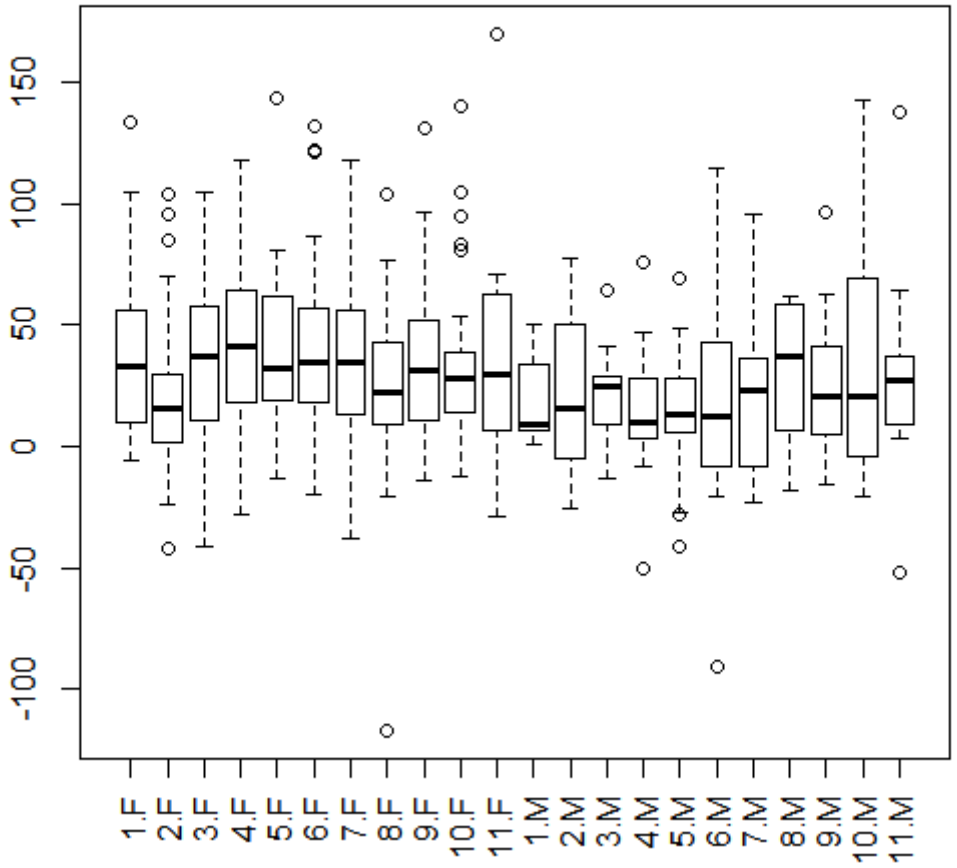

Figura 8: Valores da variação glicêmica por tempo e sexo

Com os dados obtidos e, feita a análise exploratória, começa-se então a trabalhar com a escolha dos métodos a serem utilizados, para adequação dos dados. Como objetivo também, testar os efeitos fixos, verificando se estes influem ou não na modelagem da glicemia. Além disso, serão estudados os métodos de diagnósticos, para

então, analisar os modelos gerados, obtendo quais deles possuem ajuste aos dados.

3.4 Ajuste do modelo

O interesse neste estudo é buscar alguma explicação sobre os níveis de glicemia, incorporando no modelo a variabilidade existente de cada indivíduo participante. Partindo ,então, à um modelo linear misto. Apesar de os resultados sobre o modelo serem obtidos apenas no ajuste, pode-se, através da análise descritiva, notar que a curva da variação glicêmica inicia-se em vários pontos distintos, sugerindo efeito da variabilidade do indivíduo no intercepto, optando-se por ajustar o modelo com intercepto aleatório.

A partir da definição do efeito aleatório como o intercepto, passa-se a buscar quais serão os efeitos fixos pertinentes ao modelo, para isso foi estabelecido o método de estimação da Máxima Verossimilhança (MV), e com isso foi feita a identificação dos efeitos fixos significativos.

Ao escolher o efeito aleatório e o efeito fixo, passa-se ao próximo passo - escolha da melhor estrutura de covariâncias intra-indivíduos, R_i. Para esta seleção foi usado o método de estimação MV ou MRV (máxima verossimilhança restrita), Teste da Razão de Verossimilhança e os valores AIC e BIC. As estruturas que foram selecionadas para serem testadas no modelo foram: UN, VC e CS que estão disponíveis no R. Excluindo-se de princípio as estruturas ARH(1), ARMA(1,1),AR(1), tendo em vista que estas matrizes exigem intervalos de tempo iguais entre as observações, o que não é o caso dos dados a serem trabalhados.

Inicia-se o ajuste dos dados, observando-se as matrizes de correlação e covariância

entre as medidas tomadas em cada aula, que é dada a seguir. Os valores das covariâncias são dados da diagonal principal para cima, e a correlação é representada pela parte inferior restante.

$$\begin{pmatrix}
997.03 & 284.05 & 396.67 & 427.87 & 491.01 & 612.22 & 737.09 & 109.89 & 327.73 & 676.42 & 158.65 \\
0.25 & 1215.75 & 544.86 & 364.12 & 308.21 & 790.06 & 204.64 & 294.54 & 287.40 & 219.35 & 229.90 \\
0.38 & 0.48 & 1041.63 & 444.12 & 462.06 & 740.95 & 513.54 & 604.05 & 390.55 & 380.44 & -55.85 \\
0.38 & 0.30 & 0.39 & 1211.45 & 385.02 & 768.17 & 430.94 & 52.84 & 385.25 & -6.73 & 308.53 \\
0.44 & 0.25 & 0.40 & 0.31 & 1219.72 & 615.76 & 336.15 & -10.83 & 356.74 & 542.84 & -44.91 \\
0.43 & 0.50 & 0.51 & 0.49 & 0.39 & 1992.74 & 599.78 & 645.98 & 215.59 & 774.53 & 581.51 \\
0.64 & 0.16 & 0.43 & 0.34 & 0.26 & 0.37 & 1310.58 & 303.43 & 291.93 & 520.47 & 254.89 \\
0.09 & 0.22 & 0.49 & 0.04 & -0.008 & 0.38 & 0.22 & 1413.91 & 197.49 & 482.94 & -223.65 \\
0.31 & 0.24 & 0.36 & 0.33 & 0.30 & 0.14 & 0.24 & 0.15 & 1114.64 & 429.48 & 82.91 \\
0.50 & 0.14 & 0.27 & -0.004 & 0.36 & 0.41 & 0.34 & 0.30 & 0.30 & 1772.80 & 440.51 \\
0.12 & 0.15 & -0.04 & 0.21 & -0.03 & 0.31 & 0.17 & -0.14 & 0.06 & 0.25 & 1709.01
\end{pmatrix}$$

Assim, com a constatação da correlação intra-indivíduos, parte-se, para de fato, selecionar o melhor modelo, em relaçao aos seus efeitos fixos. Dessa forma, foram feitos os testes e analisados os critérios de informação definidos anteriormente. Partindo, então do modelo completo, ou seja, modelo com todos os efeitos fixos disponíveis. E após isso, retira-se o efeito que menos significa para o modelo, fazendo-se uma nova verificação, até que restem apenas os feitos fixos significativos.

O processo de seleção será apresentado a seguir:

O modelo f1 é o completo, possui tempo, sexo, insulina, idade, diagnóstico e cidade como efeitos fixos. Percebe-se na seguinte saída do R que dentro da variável cidade há algumas com baixa significação para o modelo. Então foi feita eliminação das cidades que não eram signficantes, gerando o segundo modelo, f2. Logo, no

modelo f3, foi desconsiderado o feito fixo diagnóstico, pois apresentava a pior significância dentro do modelo. No quarto modelo, f4 retirou-se o feito tempo, agora, sexo foi a que menos contribuiu no modelo, ou seja, o modelo f5 conta apenas com idade, insulina e cidade. No sexto modelo, f6, retirou-se cidade dos efeitos fixos, por último o modelo f7 retira a variável insulina, tendo em vista que também não foi significativa, como pode ser visto, apenas a variável idade é significativa ao modelo.

```
Fixed effects: glicemia ~ tempo + sexo + insulina + idade + diagnostico + cidades
                        Value Std.Error  DF    t-value p-value
(Intercept)         1160.7082 1036.4167 379  1.1199243  0.2635
tempo                  0.6177    0.4845 379  1.2749563  0.2031
sexoM                 -9.5016    6.9895  28 -1.3594155  0.1849
insulinaS              7.4168   10.0648  28  0.7369115  0.4673
idade                  0.6191    0.2727  28  2.2705984  0.0311
diagnostico           -0.5903    0.5171  28 -1.1415552  0.2633
cidadesCEILANDIA      38.8111   20.0773  28  1.9330797  0.0634
cidadesPLANALTINA    -15.3947   19.7910  28 -0.7778648  0.4432
cidadesSAMAMBAIA      21.4709   10.5061  28  2.0436591  0.0505
cidadesSAOSEBASTIAO   13.7148   22.8782  28  0.5994711  0.5537
cidadesVALPARAISO      9.4814    9.0057  28  1.0528227  0.3014
```

Figura 9: Modelo f1

```
Fixed effects: glicemia ~ tempo + sexo + insulina + idade + diagnostico + cidade
                   Value Std.Error  DF    t-value p-value
(Intercept)     529.0643  887.9690 379  0.5958139  0.5517
tempo             0.6177    0.4845 379  1.2749561  0.2031
sexoM            -9.6709    6.7545  31 -1.4317709  0.1622
insulinaS        12.9101    8.6322  31  1.4955745  0.1449
idade             0.7183    0.2517  31  2.8542933  0.0076
diagnostico      -0.2748    0.4422  31 -0.6215513  0.5388
cidadeCEILANDIA  33.4450   19.1181  31  1.7493864  0.0901
cidadeSAMAMBAIA  15.6244    8.4113  31  1.8575545  0.0728
```

Figura 10: Modelo f2

```
Fixed effects: glicemia ~ tempo + sexo + insulina + idade + cidade
                    Value Std.Error  DF   t-value p-value
(Intercept)     -22.75224 16.887234 379 -1.347304  0.1787
tempo             0.61770  0.484490 379  1.274956  0.2031
sexoM            -9.43448  6.678797  32 -1.412601  0.1674
insulinaS        15.06764  7.827372  32  1.924994  0.0632
idade             0.74759  0.244835  32  3.053445  0.0045
cidadeCEILANDIA  33.66466 18.930673  32  1.778313  0.0849
cidadeSAMAMBAIA  15.17895  8.299929  32  1.828805  0.0768
```

Figura 11: Modelo f3

```
Fixed effects: glicemia ~ sexo + insulina + idade + cidade
                    Value Std.Error  DF   t-value p-value
(Intercept)     -19.04602 16.635155 380 -1.144926  0.2530
sexoM            -9.43448  6.678797  32 -1.412601  0.1674
insulinaS        15.06764  7.827372  32  1.924994  0.0632
idade             0.74759  0.244835  32  3.053445  0.0045
cidadeCEILANDIA  33.66466 18.930673  32  1.778313  0.0849
cidadeSAMAMBAIA  15.17895  8.299929  32  1.828805  0.0768
```

Figura 12: Modelo f4

```
Fixed effects: glicemia ~ insulina + idade + cidade
                     Value Std.Error  DF   t-value p-value
(Intercept)     -27.566184 15.735273 380 -1.751872  0.0806
insulinaS        16.927710  7.831326  33  2.161538  0.0380
idade             0.832297  0.240932  33  3.454493  0.0015
cidadeCEILANDIA  27.075009 18.621563  33  1.453960  0.1554
cidadeSAMAMBAIA  13.955953  8.378223  33  1.665741  0.1052
```

Figura 13: Modelo f5

```
Fixed effects: glicemia ~ insulina + idade
                   Value Std.Error  DF   t-value p-value
(Intercept) -22.118944 15.889051 380 -1.392087  0.1647
insulinaS    13.351384  7.896520  35  1.690793  0.0998
idade         0.802465  0.247289  35  3.245044  0.0026
```

Figura 14: Modelo f6

```
Fixed effects: glicemia ~ idade
                   Value Std.Error  DF   t-value p-value
(Intercept) -19.61025 16.222888 380 -1.208801  0.2275
idade         0.80168  0.253593  36  3.161292  0.0032
```

Feita a seleção dos efeitos fixos e aleatórios do modelo, fica restando selecionar a matriz de correlação que mais representa os dados, para isso, e baseando-se na análise descritiva feita anteriormente, usa-se apenas as matrizes VC, UN, CS - componente de variância, não estruturada e simetria composta, respectivamente. Tendo em vista que a seleção do modelo fixo resultou no f7. Será utilizado suas variações para cada matriz, onde pode-se verificar as relações no quadro a baixo:

Tabela 3: Estatísticas de ajuste dos modelos

Modelo	Estimação	AIC	BIC	-2LogVeross	Estrutura R_i
f7.1	MVR	4123.65	4139.77	2057.82	VC
f7.2	MVR	4128.81	4148.99	2059.40	CS
f7.3	MVR	4152.49	4390.58	2017.24	UN

Com a observação da tabela, nota-se que para os valores de AIC, BIC e MV, a matriz que mais se ajusta ao modelo é a de VC (componente de variância), pois é o que apresenta menor valor.

Agora, é preciso fazer uma análise de diagnóstico, para assim, verificar se o modelo está totalmente adequado aos dados. A análise será feita a partir da verificação dos gráficos que se seguem:

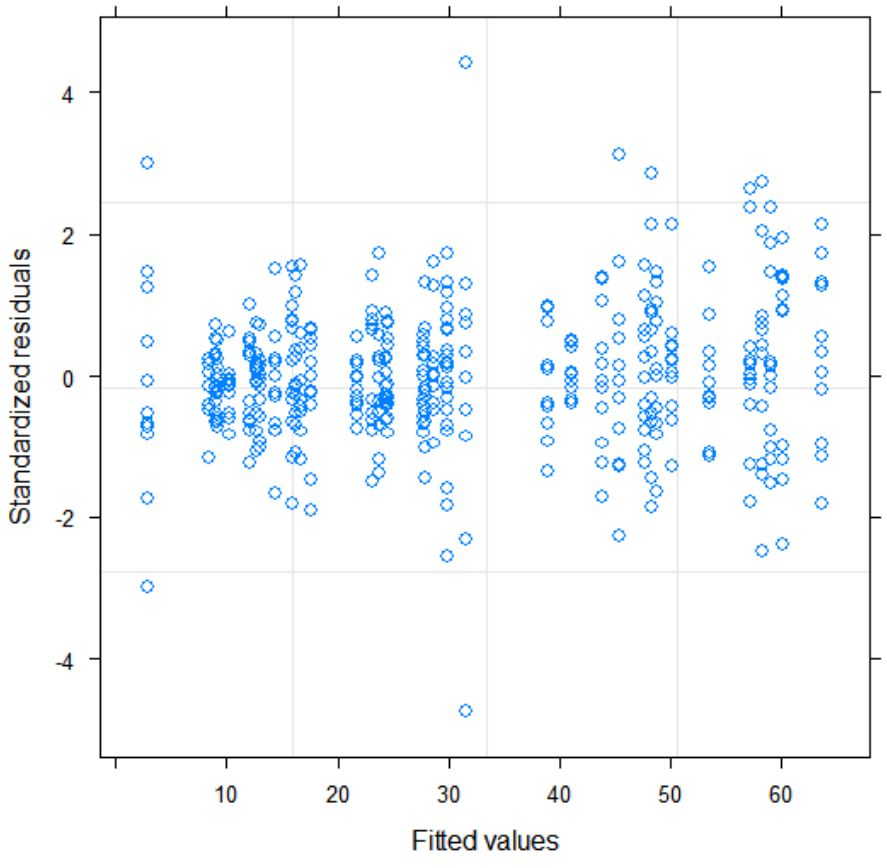

Figura 15: Resíduos padronizados

No primeiro gráfico foi obtida a padronização dos resíduos. A forma pela qual os pontos se espalham pelo gráfico é aleatória, ou seja, há a indicação de aleatoriedade nos resíduos do modelo. Por último, o gráfico, abaixo, da normalização dos resíduos, mostra que há a verificação de os erros possuirem distribuição normal. Tendo em vista o diagnóstico feito ao modelo, com a visualização gráfica, nota-se que há um bom ajuste do modelo aos dados estudados.

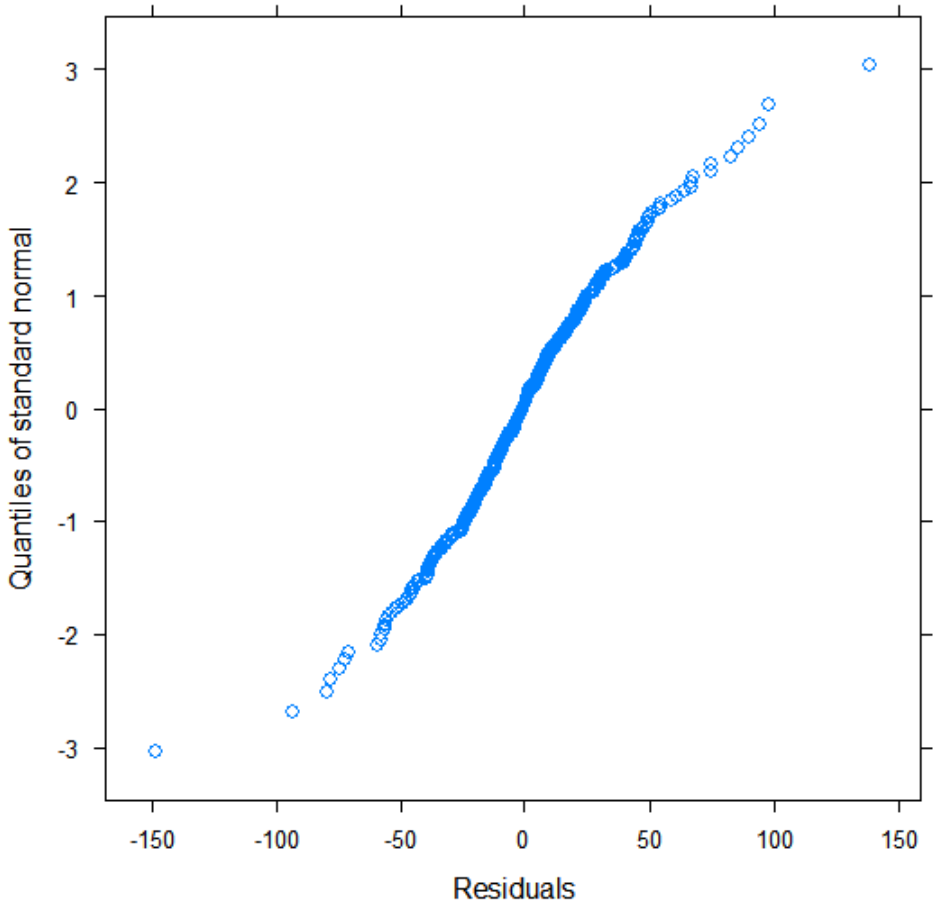

Figura 16: Resíduos normalizados

4 Conclusões

Este trabalho teve como foco os níveis de glicemia dos participantes do Projeto Doce DESAFIO. Neste contexto, o problema foi abordado por meio do modelo linear misto, estudando sua adequacidade através da análise de resíduos.

O modelo linear misto mostrou-se bastante eficaz no ajuste à variação dos níveis de glicemia dos participantes. Dentre os modelos analisados, o modelo com intercepto aleatório, e matriz de componentes de variâncias - VC, juntamente com o efeito fixo idade, foi o que se apresentou mais adequado. A análise de resíduos não evidenciou afastamento da suposição de resíduos aleatótios, bem como, da verificação de normalidade do erro.

A função lme() do *software* R mostrou-se um bom pacote para análise de modelos lineares mistos, possuindo fácil implentação, e grande abrangência. Sendo assim, o *software* R foi decisivo no estudo.

Referências

BARBOSA, M. Uma abordagem para análise de dados com medidas repetidas utilizando modelos lineares mistos. 2009. Tese (Mestrado em Agronomia) - Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba 2009.

COSTA, T.R. Modelos lineares mistos: uma aplicação na produção de leite de vacas da raça sindi. 2010. Tese(Mestrado em Biometria e Estatística Aplicada), Universidade Federal Rural de Pernambuco, Recife 2010.

DAVIDIAN, M. Applied Longitudinal Data Analysis, Spring 2007. Disponível em:¡http://www.stat.ncsu.edu/people/davidian/courses/st732¿. Acesso em: 08 fevereiro 2013.

EVERRIT, B.S. Na R and S-Plus companion to multivariate analysis. London: Springer-Verlag, 2005. 221 p.

FARAWAY, J.J, Extending the Linear Model with R Generalized Linear, Mixed Effects and Nonparametric Regression Models. Taylor and Francis Group, LLC, 2006.

Global Health Observatory Data Repository. Organização Mundial da Saúde (OMS). Disponível em:¡http://apps.who.int/gho/data/view.main.2480?lang=en¿. Acesso em: 23 julho 2013 LAIRD, N.M.; WARE ,J.H. Random effects models for longitudinal data. Biometrics, Washington, 1982.

LIANG, Kung – Yee; ZEGER, Scoot L. Longitudinal analysis using generalized linear models. Biometrika 73 p. 1986.

LIMA, C.G. Análise de dados longitudinais provenientes de experimentos em blocos caualizados. 1996 119 p. Tese (Doutorado em Estatística e Experimentação Agronômica) – Escola Superior de Agricultura “Luiz de Queiroz”, Universidade de São Paulo, Piracicaba, 1996.

NATIS, I. Modelos lineares hierárquicos. 2000. 87 p. Dissertação (Mestrado em Estatística) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 2000.

NOBRE, J. S. Métodos de diagnostico para modelos lineares mistos. Dissertação de mestrado, IME/USP, São Paulo, 2004.

PINHEIRO, J.C. Topics in mixed effectsmodels. 1994. 210 p. Thesis(PhD) – University of Wisconsin, Madison, 1994.

PINHEIRO, J. C.; BATES, D. M. Mixed-effects models in S and S-PLUS. New York: Springer - Verlag, 2000, 528p

VERBEKE, G.; MOLENBERGHS, G. Linear mixed models for longitudinal data. New York: Springer - Verlag, 2000, 568p.